中華農業研究 (Jour. Agric. Res. China) 49(2): 36～48 (2000)

離群值鑑定方法應用於作物區域試驗穩定性分析之例釋[1]

呂秀英[2]

摘要：本省作物區域試驗一般採用的穩定性介量多以最受廣泛使用的直線迴歸分析法爲主。由於傳統迴歸分析是以最小平方法(LS)來估計迴歸係數，對離群值(或異常值)之存在非常敏感，離群值之存在會造成迴歸係數估計不良，也會造成直線迴歸分析殘差均方異質，從而可能影響穩定性分析之準確度。現行用以診斷離群值的統計方法甚多，可區分爲兩大類別：傳統的迴歸診斷及最新發展的穩健迴歸。然而，它們卻從未被引用到作物區域試驗資料之穩定性分析。爲尋求適當可行的方法，對作物區域試驗資料作一鑑定與判斷，以找出有影響力的離群值，本研究利用 Yates and Cochran (1938) 之大麥資料作爲材料，進行各種離群值診斷法在穩定性分析的應用分析。結果得知，以LS殘差爲基礎的傳統迴歸診斷法並不適合作爲離群值診斷工具，而利用穩健迴歸可有效地辨識出資料的型式及偵測出離群值。

關鍵詞：區域試驗、穩定性、離群值診斷、穩健迴歸、再加權最小平方迴歸。

前　言

育種程序中區域試驗施行之目的，在選出高產且穩定的作物品種。本省作物區域試驗一般採用的穩定性介量多以最受廣泛使用的直線迴歸分析法爲主。直線迴歸模式能有效地估計數量性狀對環境的適應能力，其所估得之迴歸係數並有容易解釋的優點，然而，由於其受限於某些統計前提，故僅適用於以下場合[5,14,15,26,29,36,41,42]：(1)基因型、環境及重複數很多；(2)基因型與環境交感效應大部份可用直線關係來解釋；(3)基因型之迴歸殘差均方之間互爲均質；(4)基因型之機差變方均質(如一般變方分析所需之假設前提)。當這些條件中任何一個不符合時，所求得之穩定性介量的估計值便不足以代表基因型對環境之適應能力。

區域試驗資料的準確度，將嚴重影響分析結果的可信賴程度。進行迴歸分析的資料集倘含有某些觀測值，與其他觀測值所表示之型態比較呈不正常者，稱之爲離群值(outlier)。由於傳統迴歸分析是以最小平方法(least squares method, LS)來估計迴歸係數，此法對離群值之存在非常敏感，離群值之存在會造成迴歸係數估計不良，進而影響模式變數之選擇[8]，故可能會對穩定性迴歸介量之估計造成很大的偏差。另外Lu[28]亦發現，離群值之存在也是造成穩定性直線迴歸分析中殘差均方異質的主要因素之一，而偵測及妥當處理離群值要比進行資料轉換更能解決均方異質的問題。

關於離群值鑑定法之特殊技術日異又新，各具理論基礎及適用場合，且效率不一，但基本上其所

---

1. 行政院農業委員會農業試驗所　研究報告第2023號。

2. 本所農藝系研究員。臺灣省　臺中縣　霧峰鄉。

偵出之離群值僅代表數理上的不正常值，一旦經檢測出來，未必就須將之立即摒棄，而尙應就其是否因人爲疏忽或另有生物學上之因素，探明其發生的背景及來源，而後才能決定處理離群值的方式[38]，例如隨後修正或刪除、從中進一步發掘出其它被漠視之重要且新的訊息、藉由模式之矯正或估計方法予以合併、認知資料中伴隨著的問題而作更進一步的試驗等[6,12]。因此，若欲將離群值鑑定之技術應用於穩定性分析上，尙須配合實際作物區域試驗的資料架構，才不致將基因型對環境之確實極端表現誤認爲是不正常的離群值。有關穩定性分析之離群值鑑定等相關課題，迄今仍少有準則。在進行穩定性分析之前，倘若能尋求出適當可行的診斷方法，對參試資料作一鑑定與判斷以找出眞正的離群值，當更能提高穩定性分析之準確度。因此本研究利用 Yates and Cochran[43]之大麥區域試驗資料爲材料，進行各種離群值診斷法的應用分析，以探討其應用在穩定性分析的可行性。

統計方法

傳統線性模式如下：

$$Y_{n\times r} = X_{n\times P}\,\theta_{p\times 1} + \varepsilon_{n\times 1} \quad \cdots\cdots(1)$$

即

$$y_i = x_{i1}\,\theta_1 + \ldots + x_{ip}\,\theta_p + e_i \quad \cdots\cdots(2)$$

，i=1,...,n ，j=1,...,p ，式中 y_i 爲隨變數(response variable)， $x_{i1},\ldots,x_{ip}$ 爲自變數(explanatory variables)， e_i 爲機差項而滿足平均值爲 0 且變方爲 σ^2 的常態分布， θ_j 爲未知介量(unknown parameters)，即迴歸係數(regression coefficient ，包括截距及斜率)。設殘差(residual) r_i 爲 y_i 之實測值與其預測值之差，即

$$r_i = y_i - \hat{y}_i \quad \cdots\cdots(3)$$

則一般皆以最小平方法(least squares method ， LS)來估計迴歸係數，即

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} r_i^2 \quad \cdots\cdots(4)$$

今假設存在有離群值 x_1，由於 x_1 離開其他 $x_2, \ldots x_n$ 值很遠，故離原來直線之殘差 r_1 變得很大，而導致 $\Sigma\, r_i^2$ 也變大，在 LS 之定義下，爲使 r_i^2 變小及 $\Sigma\, r_i^2$ 最小，必須重新配合成新的直線，因此不免會使 $r_2^2,\ldots,r_n^2$ 多少增大，所以 LS 對離群值之存在非常敏感，離群值之存在會造成迴歸係數估計不良[32]。

當自變數爲固定值時，y 方向的離群值(又名垂直離群值， vertical outlier)可簡單地由殘差值或殘差圖被發現出來；但如自變數爲逢機變異的觀測值時，x 方向的偏離也可能發生，一般將之稱爲槓桿點(leverage point)。只在 x 方向偏離的槓桿點，未必就是迴歸離群值，當其接近於大部份資料所決定的迴歸線時，稱爲「好」的槓桿點，因爲 x_i 雖遠離但 y_i 卻配合直線模式，故可改進迴歸係數之準確度，即表示該槓桿點對迴歸係數雖有強烈影響的潛力，但不見得會表現出來；反之，倘槓桿點遠離於大多數資料座落之平面，則稱此點爲「壞」的槓桿點，爲最有害的離群值，因爲它會吸引甚至使傳統 LS 迴歸傾斜（此在觀念上與機械槓桿作用相類似，故稱之爲槓桿），即表示該槓桿點對迴歸分析有很大的影響。爲區分槓桿點的好壞，必須同時考慮 y_i 及 x_i [33]。

所謂離群值診斷法爲利用統計量來偵測出對 LS 估值有大影響的觀測值。一般解決迴歸離群值之統計方法，目前大致可有兩種解決途徑，茲將其整理摘錄如表 1 所列，並詳細說明如下：

(一) 迴歸診斷 (regression diagnostics)

由於傳統迴歸分析係以最小平方法(LS)估計迴歸係數，故傳統的離群值診斷法都是利用 LS 之殘差分析(residual analysis)來偵測離群值。這類用來判斷迴歸影響點的殘差統計量種類甚多，原則上針對單

表 1. 迴歸離群值鑑定的統計量摘錄

Table 1. A summary of statistics for the regression outlier detection

Diagnostics	Symbol	Definition	Cut-off value	Author(user)
〔Regression diagnostic〕				
Diagonal elements of hat matrix	h_{ii}	$H = X(X^tX)^{-1}X^t$	2p/n	
Mahalanobis distance	MD_i^2	$(n-1)\times(h_{ii} - 1/n)$	$\chi^2_{p-1,\alpha}$	
Standarized residual	r_i/s	$r_i/s\sqrt{1 - h_{ii}}$	2.5	
Studentized residual	t_i	$r_i/s\sqrt{1 - h_{ii}}$	2.5	
Jackknifed residual	t(i) (RSTUDENT)	$r_i/s(i)$	2.5	
Cook's squared distance	$CD^2(i)$	$\frac{t_i^2}{p} \times \frac{h_{ii}}{1 - h_{ii}}$	1.0	Cook (1977)
DFFITS(i)	DFFITS(i)	$\frac{r_i}{s(i)} \times \frac{\sqrt{h_{ii}}}{1 - h_{ii}}$	$2\sqrt{p/n}$	Belsley *et al.* (1980)
$DFBETAS_j(i)$	$DFBETAS_j(i)$	$\frac{c_{ji}}{\sqrt{\sum_{k=1}^{n} c_{jk}^2}} \times \frac{r_i}{s(i)(1 - h_{ii})}$	$2\sqrt{n}$	Belsley *et al.* (1980)
Weighted regression	WLS (Ws,Wh)	fit from smooth or hard weighted function		Holland and Welsch (1977); Hogg (1979)
〔Robust regression〕				
Robust residual	r_i/σ^*	standardized LMS residuals	2.5	Rousseeuw (1984); Atkinson (1986)
Reweighted least squares	RLS(W)	robust fit from weighted function based on LMS		Rousseeuw and Leroy (1987)
Robust distance	RD_i	RD_i based on MVE	$\sqrt{\chi^2_{p,1-\alpha/2}}$	Rousseeuw and Zomeren (1990)

Linear model is $Y_{n\times 1} = X_{n\times p}\,\theta_{p\times 1} + \varepsilon_{n\times 1}$, n=sample size, p=number of explanatory variables.

s, s(i), c_{ji}, σ^* defined as text.

一離群值的鑑定方法，其效率多已獲得證實，但對於多個以上之離群值則不易被檢測出來，且涉及之計算均頗爲複雜，故較少被實際採用[32]。

在傳統診斷工具中最常被使用到的統計量爲LS預測空間的投影矩陣(projection matrix, H)的對角線元素(h_{ii})；$H = X(X^tX)^{-1}X^t$，因爲 $\hat{y} = Hy$ 是將觀測值y轉變成LS估值 $\hat{y}$，故一般也稱H爲銜接矩陣(hat matrix)。大多數學者[12,16,17,18,37]以 $h_{ii}>2p/n$ 來決定有潛力的影響點，此處的2p/n謂爲切斷值(cut-off value)。由於H僅以x_i爲基礎，故只能指出x方向的離群值，至於y方向之離群值則完全被忽略了。另外，由於LS在定義上爲嘗試要避開較大殘差，往往會犧牲其他觀測值，因此一個離群值即可能會造成

多數資料的配合不良，於是該離群值便產生較小的LS殘差，反而無法被發掘出來，特別是在該離群值爲槓桿點之場合，換言之，以LS殘差爲基礎的h_{ii}診斷法容易受到遮蔽效應(mask effect)的影響，使得有時候一些眞正的離群值會被好的槓桿點之效果所遮掩住，故僅適用於在x方向之單一離群值的情形。其他診斷法只要是涉及到H之使用，某些槓桿點便可能會被其他所隱蔽[32]。

在其他診斷迴歸離群值的場合中，也有利用馬哈拉諾比斯距離(Mahalanobis distance, MD_i^2)、標準化殘差(standarized residuals, r_i/s)、學生氏化殘差(studentized residuals, t_i)、摺刀化殘差(jackknifed residuals, t(i))等統計量來偵測迴歸的槓桿效應。MD_i^2與h_{ii}之間具有一對一之關係，且與$\chi^2_{(p-1)}$比較可判斷第i項與一堆資料間遠離的程度。r_i/s式中的s爲s^2之開方值，$s^2=\sum_{i=1}^{n} r_i^2/(n-p)$爲$\sigma^2$之無偏估值。$t_i$與t(i)最大的不同在於後者爲除去第i項後重新再進行迴歸分析的學生氏化殘差，其名稱不一，如Velleman and Welsch[39]及Cook and Weisberg [12]稱t_i爲內部學生氏化殘差(internally studentized residual)，而t(i)爲外部學生氏化殘差(externally studentized residual)，但是Belsley *et al.* [7]以"RSTUDENT"來表示t(i)，Atkinson[2]則指t(i)爲交叉驗證殘差(cross-valiadatory residual)。r_i/s、t_i、t(i) 一般皆以2.5作爲切斷點。

另外，爲檢查第i個觀測值的影響程度，可利用分別含或不含該觀測值進行迴歸分析，基此產生庫克氏平方距離(Cook's squared distance, $CD^2(i)$)[11]之統計量；$CD^2(i)$與完整資料的三個統計量有關 (公式請參見表1)：迴歸係數個數(p)、第i個觀測值的學生氏殘差(t_i)、y_i變方($\sigma^2 h_{ii}$)對e_i變方($\sigma^2(1-h_{ii})$)之比值，因此$CD^2(i)$由兩個成分組成，一個成分反映模式配合y_i之程度，另一個則是測量x_i遠離其他x_k有多遠，兩者同時促成$CD^2(i)$的值變大。而 Belsley *et al.* [7]所提出的DFFITS(i)與$CD^2(i)$十分相似，兩者唯一主要差異在於DFFITS(i)是以s(i)取代s，故當刪去一個觀測值時可用以測量其對預測之影響。Belsley *et* al. [7]又以第j迴歸係數上之改變爲基礎，定義了另一種診斷統計量：$DFBETAS_j(i)$，其計算公式中的$C=(X^tX)^{-1}X^t$，稱爲捕手矩陣(catcher matrix)。一旦觀測值之$CD^2(i)>1.0$，或$DFFITS(i)>2\sqrt{p/n}$，或$DFBETAS_j(i)>2\sqrt{n}$，則應該詳加檢查。

t(i)、$CD^2(i)$、DFFITS(i)與$DFBETAS_j(i)$這一類的診斷都是以一次刪除一項爲原則，比較刪去第i項後所計算結果之差異來判別其對迴歸分析的影響，此方式統稱爲單項診斷(single-case diagnostics)。至於將其擴充至多項診斷(multiple-case diagnostics)的情形，雖然也有學者[1,12,23,27]陸續提出不同方法，但由於計算時間上之限制，且誰也不敢保證能將資料中存在的所有不好的資料群完全確認出來，故並不如單項診斷法之廣受使用。因此，本研究不擬採用這些目前尙乏人問津的多項診斷法。

由於LS容易受到離群值之影響其迴歸估計值，因此另有些學者[19,20,40]提出加權迴歸分析法(weighted regression analysis, WLS)，就彼此間殘差之關係加權，針對具有較大殘差之資料給予較低之加權值，藉此以獲致較準確的估計值，即

$$b_w=(X^tWX)^{-1}X^tWy \qquad \text{.......(5)}$$

W值在0及1之間，可以利用一個$|r_i/s|$的非遞增函數來決定第i個觀測值的加權，例如

$$Wh_i=\begin{cases}1 & \text{if } |r_i/s|\le 2.5\\ 0 & \text{if } |r_i/s|>2.5\end{cases} \qquad \text{.......(6)}$$

，此爲一種強硬拒絕(hard rejection)型態的加權函數，在接受與拒絕點之間產生徹底的二分法；另外也可採取緩和拒絕(smooth rejection)的加權函數，定義如下[21]：

$$Ws_i=\begin{cases}1 & \text{if } |r_i/s|\le k\\ k|s/r_i| & \text{if } |r_i/s|>k\end{cases} \qquad \text{.......(7)}$$

理論上k值通常選擇約1.5[20]。一般而言，由於此法仍是以LS殘差爲基礎，因此很難保證加權後能產生

一個非常穩健的估計值，因此WLS之應用也只是將產生的Wh或Ws視爲離群值診斷的另一種統計量而已，亦即倘若觀測值之加權值小於0.5，則須進一步檢查是否因人爲疏忽所引起或資料來自於不同族群[8]。

傳統LS迴歸診斷分析在目前很多統計套裝軟體，爲例行的程序。因此本研究有關各種以LS殘差爲基礎的迴歸診斷統計量之計算，均係依據SAS(Statistical Analysis System)[35]統計軟體，在個人電腦上建檔及運算。

(二)穩健迴歸 (robust regression)

由於配合LS所產生的殘差會受到遮蔽效應之影響而無法應付多離群值的情形，因此最適合作爲診斷工具的殘差必須由一個非常穩健之估計值產生，才能將資料中所有離群值檢視出來。針對此問題，Rousseeuw[30]將LS法(眞正全名爲 least sum of squares)中平方的和(sum)以中量(median)取代，於是提出了最小平方中量法 (least median of squares method, LMS)，即

$$\underset{\hat{\theta}}{\text{Minimize}} \ \underset{i}{\text{med}} \ r_i^2 \quad \cdots\cdots (8)$$

此估計值同時對y及x的離群值均十分穩健，而且由於LMS對離群值之抵抗能力與p無關，因此在多變數的場合，爲相當可靠的資料分析工具。Atkinson [3]並建議在使用無向度(dimension-less)的標準化LMS殘差(r_i/σ^*)來偵測出離群值後，應再以LS作爲診斷，最後藉由LS及LMS兩種分析結果之差異性比較，來確認離群值存在的事實；另外，標準化LMS殘差也可用來作爲推論隨變數是否需轉換的診斷方法[4,10]。標準化LMS殘差又稱爲穩健殘差(robust residual)[33]。

爲使LMS能有如同t值及信賴界限等之標準統計量，可利用離群值診斷爲基礎，導入加權的觀念，例如可使用以下加權：

$$W_i = \begin{cases} 1 & \text{if } |r_i/\sigma^*| \le 2.5 \\ 0 & \text{if } |r_i/\sigma^*| > 2.5 \end{cases} \quad \cdots\cdots (9)$$

式中

$$\sigma^* = 1.4826(1+\frac{5}{n-p})\sqrt{\underset{i}{\text{med}}\ r_i^2} \quad \cdots\cdots (10)$$

此意謂著如果標準化LMS殘差不大則保留在加權LS中，倘一旦確實是個離群值，則無視它的存在，即

$$\underset{\hat{\theta}}{\text{Minimize}} \sum_{i=1}^{n} W_i r_i^2 \quad \cdots\cdots (11)$$

此法稱爲再加權LS (reweighted least squares, RLS)[32]。對任何檢定等級而言，產生的W_i也都可以視爲是離群值診斷。

標準化LMS殘差(r_i/σ^*)可用來指出迴歸離群值，但無法指出迴歸離群值中何者爲槓桿點，基此，Rousseeuw and Zomeren[33]利用最小體積橢圓體估值(minimum volume ellipsoid estimator, MVE) [31]，以涵蓋半數觀測值之MVE中心爲資料集X的代數平均値，並由同一橢圓體來決定樣本共變方矩陣，而定義出另一個穩健的診斷統計量，稱爲強建距離(robust distance, RD_i)，其切斷點爲$\sqrt{\chi^2_{p,1-\alpha/2}}$。此穩健共變方矩陣在許多不同型態的多變量分析中均可用來鑑定離群值，例如主成份分析[9,13]及典型相關分析[24]等。RD_i在計算上需花費的時間雖然較多，但使用者可以很容易而立即獲致更多的訊息。

但因爲RD_i中未使用到y_i，若僅憑RD_i亦不能區分槓桿點之好壞，因此 Rousseeuw[34]建立一個新的表達方式，以穩健殘差(r_i/σ^*)對穩健距離(RD_i)作圖，可說明資料的各種型式，以簡單迴歸爲例，如圖1所示，以$\sqrt{\chi^2_{p,1-\alpha/2}}$ 爲垂直分界線，其右邊即爲槓桿點，落在 [-2.5, 2.5] 之水平容忍度之外則爲迴歸離

群值，因此圖中(a)爲大部份的正常資料點，(b)及(d)點爲偏離線性型式的離群值，(c)則非離群值；(c)，(d)均爲槓桿點，因其 x_i 值偏離，因此稱(c)爲好槓桿點，(d)爲壞槓桿點，而(b)爲垂直離群值，即爲迴歸離群值但非槓桿點。綜言之，從 r_i/σ^* 對 RD_i 之關係圖中可將資料區分成以下四種類型：

(a) 正　常　值　：$|r_i/\sigma^*| \leqq 2.5$，$RD_i \leqq \sqrt{\chi^2_{p,1-\alpha/2}}$

(b) 垂直離群值　：$|r_i/\sigma^*| > 2.5$，$RD_i \leqq \sqrt{\chi^2_{p,1-\alpha/2}}$

(c) 好槓桿點　：$|r_i/\sigma^*| \leqq 2.5$，$RD_i > \sqrt{\chi^2_{p,1-\alpha/2}}$

(d) 壞槓桿點　：$|r_i/\sigma^*| > 2.5$，$RD_i > \sqrt{\chi^2_{p,1-\alpha/2}}$

另外，由於數理上的切斷點多少有些武斷，因此由穩建殘差與穩健距離之圖示中，尚可根據觀測點遠離分界線的程度，來辨識出所謂的邊界情形(boundary case)，亦即當觀測點在X空間上不致偏離太遠時，通常可將其視爲一個溫和的迴歸離群值[34]。

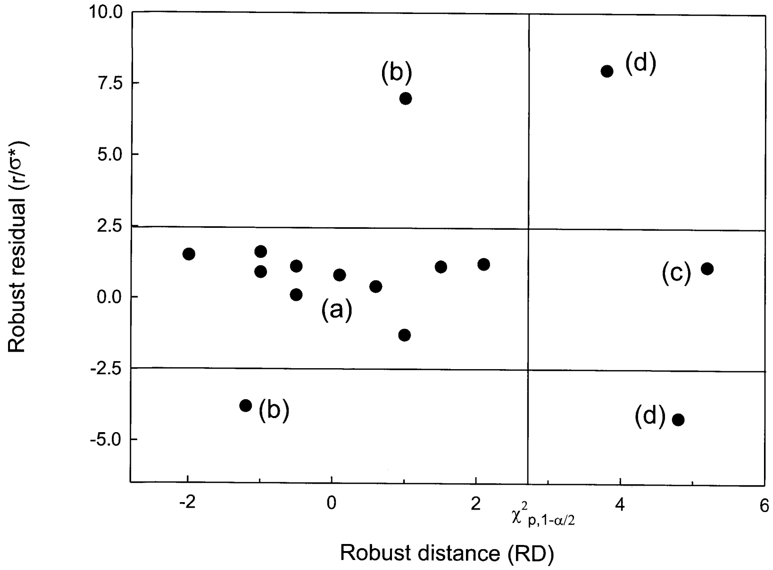

圖 1. 以簡單迴歸爲例，根據穩健殘差對穩健距離之關係圖，顯示出(a)正常值、(b)垂直離群值、(c)好的槓桿點、(d)壞的槓桿點的位置所在。

Fig. 1. Simple regression example with (a) regular observations, (b) vertical outliers, (c) good leverage point, and (d) bad leverage point obtained from a diagnostic plot of robust residuals versus robust distances.

基本上，所謂穩健性 (robustness)，是想辦法使估計值不會受到離群值太大的影響，但並非如一般人所誤以爲其目的純粹在忽略離群值，相反地，它是從一種穩健或具抵抗能力的迴歸中檢視殘差以判別離群值，這是LS殘差所無法做到的。因此，迴歸診斷和穩健迴歸事實上是以相反的次序達到相同之目的：迴歸診斷是先嘗試刪除離群值後再以LS迴歸配合剩餘的正常資料，而穩健迴歸則首先希望對大多數資料配合一迴歸後，從穩健的解決方式中發現具最大殘差的離群值[32]。換言之，穩健迴歸雖較傳統的迴歸診斷之分析程序複雜，但在異常資料不慎加入分析之後，能配合迴歸模式提出一適當的矯正程序，而無視於離群值之存在 (但並非忽略)，以提高分析的準確度。

雖然穩健迴歸最近幾年應用於迴歸分析上的技術已經成熟且相當有效，但因爲發展起步較晚，仍屬於較創新的方法，一般性的統計套裝軟體至今尚未納入，不過可以採用 Rousseeuw 博士所提供之

PROGRESS及MINVOL電腦程式，在個人電腦上執行，即可獲得迅速且準確的分析結果。PROGRESS的輸出結果步驟如下[25, 32]：

(1)作y對x之點狀分佈圖；

(2)求y、x之基本統計量，如中量、分散度、標準化觀測值、Pearson相關係數、Spearman序列相關係數；

(3)求LS迴歸估值、殘差(r_i)及標準化殘差(r_i /s)；

(4)求LMS迴歸估值、殘差(r_i)及穩健殘差(r_i / σ*)；

(5)在LMS基礎下求RLS迴歸估值。

至於穩健距離(RD_i)，則可藉由MINVOL電腦程式求得，由於RD_i係從MVE產生，故該程式之輸出結果除了RD_i以外，尚包含迴歸向量>p時所估得的各種共變方矩陣(h_{ii})及馬哈拉諾比斯距離(MD_i^2)。

由於穩定性分析是以各品種之產量隨環境指標作迴歸所得的迴歸係數來評價品種的穩定性，該環境指標是以各環境之平均收量與整個試驗的總平均值之偏差，作爲該一試驗的環境效應估值，因此在穩定性迴歸模式中的自變數x，即環境指標估值，爲觀測值y的線性組合，並非固定值，因此發生在y方向的離群值，自然也會影響x方向之數值，故必須考慮能同時對x及y兩向之離群值進行診斷的方法。本研究將針對上述各種離群值診斷法進行探討，以從中尋求出適當可行的方法，使能鑑定出穩定性迴歸分析之離群值。

試驗材料

爲探討上述各種診斷方法應用在穩定性分析的可行性，本研究利用Yates and Cochran[43]於1938年所進行之大麥(barley)區域試驗資料爲分析材料，進行各種方法間的差異比較。大麥試驗產量資料及所估得之環境指標列如表2。

表2. 5個大麥品系於6種環境下生長之產量資料及所估計之環境指標

Table 2. Yield and environmental index for the 5 barley lines grown in 6 environments

Environment	Yield for each line ^{z} (bushels/acre)					Mean	Environment index
	1	2	3	4	5		
1	161.7	187.7	200.1	196.9	182.5	185.8	-16.40
2	247.0	257.5	262.9	339.2	253.8	272.1	69.90
3	185.4	182.4	194.9	271.2	219.2	210.6	8.44
4	218.7	183.3	220.2	266.3	200.5	217.8	15.62
5	165.3	138.9	165.8	151.2	184.4	161.1	-41.06
6	154.6	143.8	146.3	193.6	190.1	165.7	-36.50
Mean	188.8	182.3	198.4	236.4	205.1	202.2	

^{z} Total of three plots over two years. Source: Yates and Cochran (1938).

結果與討論

Yates-Cochran之大麥試驗資料在以LS殘差爲基礎下，每一品種以各種迴歸診斷法所得之結果於表3。由表3結果得知，所有品種在第2個環境(地區)的h_{ii}皆大於其切斷點，表示在x方向第2個觀測值有偏離其他資料點之現象，亦即環境2之環境指標爲槓桿點，此結果可由表2中環境2之指標估值高達69.90(此意謂著環境2爲一個極高產地區)得到印證；但是若由MD_i^2，r_i/s，t_i及t(i)之結果來看，並無指

表 3. 大麥資料根據LS殘差之離群值鑑定

Table 3. Outlier diagnostics based on LS residuals for the barley data

Line	Env.	h_{ii}	MD_i^2	r_i/s	t_i	t(i)	$CD^2(i)$	DFFITS(i)	$DFBETAS_j(i)$ (0.82)		WLS	
	i	(0.67)[z]	(7.38)	(2.50)	(2.50)	(2.50)	(1.00)	(1.16)	Intercept	slope	Ws	Wh
1	1	0.20	1.41	-1.00	-1.12	-1.16	0.15	-0.58	-0.53	0.23	1	1
	2	0.74	2.88	-0.06	-0.12	-0.10	0.02	-0.17	-0.08	-0.15	1	1
	3	0.18	0.83	-0.79	-0.87	-0.84	0.08	-0.39	-0.38	-0.08	1	1
	4	0.20	2.14	1.26	1.41	1.72	0.24	0.85	0.78	0.32	1	1
	5	0.37	1.88	0.84	1.06	1.08	0.32	0.82	0.55	-0.60	1	1
	6	0.32	0.87	-0.25	-0.31	-0.27	0.02	-0.19	-0.13	0.13	1	1
2	1	0.20	3.10	1.53	1.71	2.87	0.36	1.43	1.31	-0.57	0.98	1
	2	0.74	3.13	0.45	0.88	0.85	1.12	1.44	0.68	1.27	1	1
	3	0.18	0.46	-0.58	-0.64	-0.58	0.04	-0.27	-0.26	-0.06	1	1
	4	0.20	1.44	-1.02	-1.14	-1.20	0.16	-0.59	-0.54	-0.23	1	1
	5	0.37	1.05	-0.20	-0.26	-0.22	0.02	-0.17	-0.11	0.12	1	1
	6	0.32	0.82	-0.18	-0.21	-0.19	0.01	-0.13	-0.09	0.09	1	1
3	1	0.20	1.96	1.20	1.34	1.56	0.22	0.78	0.71	-0.31	1	1
	2	0.74	2.89	-0.11	-0.22	-0.19	0.07	-0.33	-0.15	-0.29	1	1
	3	0.18	0.83	-0.80	-0.88	-0.84	0.08	-0.39	-0.38	-0.09	1	1
	4	0.20	0.44	0.49	0.55	0.49	0.04	0.24	0.22	0.09	1	1
	5	0.37	1.23	0.44	0.55	0.49	0.09	0.37	0.25	-0.28	1	1
	6	0.32	2.64	-1.22	-1.48	-1.91	0.53	-1.32	-0.95	0.92	1	1
4	1	0.20	0.79	-0.71	-0.80	-0.75	0.08	-0.37	-0.34	0.15	1	1
	2	0.74	3.22	-0.52	-1.03	-1.04	1.53	-1.77	-0.84	-1.56	1	1
	3	0.18	1.70	1.15	1.27	1.42	0.17	0.65	0.64	0.14	1	1
	4	0.20	0.23	0.26	0.29	0.25	0.01	0.12	0.12	0.05	1	1
	5	0.37	2.34	-1.04	-1.30	-1.49	0.49	-1.13	-0.76	0.83	1	1
	6	0.32	1.72	0.86	1.05	1.07	0.26	0.74	0.53	-0.51	1	1
5	1	0.20	1.59	-1.07	-1.19	-1.29	0.18	-0.64	-0.59	0.26	1	1
	2	0.74	3.18	0.49	0.97	0.96	1.35	1.62	0.77	1.43	1	1
	3	0.18	0.77	0.76	0.84	0.80	0.08	0.37	0.36	0.08	1	1
	4	0.20	1.99	-1.21	-1.35	-1.59	0.22	-0.78	-0.72	0.30	1	1
	5	0.37	1.18	0.39	0.49	0.44	0.07	0.33	0.22	0.24	1	1
	6	0.32	1.29	0.64	0.78	0.73	0.14	0.50	0.36	-0.35	1	1

[z] The cut-off value.

出x方向有資料偏離的現象，而由t(i)則發現品種2在刪除環境1之資料後對迴歸分析有影響；t(i)、$CD^2(i)$、DFFITS(i)與$DFBETAS_j(i)$雖然都是以一次刪除一項爲原則，來比較刪去第i項後所計算結果之差異以判別其對迴歸分析的影響，但結果並不盡相同，由$CD^2(i)$指出環境2之存在與否對品種2,4,5較有影響，而由DFFITS(i)與$DFBETAS_j(i)$之結果，尚發現品種2在環境1以及品種3在環境6之資料的潛

在影響。就加權迴歸而言，僅有品種2在環境1之Ws=0.98，表示在品種2的迴歸分析中，祇有環境1的觀測值有偏離該品種其他環境資料的趨勢，但因為Ws值近於1且Wh等於1，顯然此離群值對迴歸分析的影響程度應該不致太大。綜合以上結果得知，探討迴歸離群值的問題，是無法僅憑藉著一、兩個傳統的離群值診斷統計量就能偵測出來。歸納上述各種診斷法之結果，大致可獲得一個簡單而清晰的結論，即環境2之環境指標甚高於其他環境值，這是一個有強烈影響潛力的槓桿點，不過是好是壞，尚無法確知，而品種2在環境1的資料可能是一個對直線配合度影響不大的離群值。由於傳統的離群值診斷法都是利用LS殘差來偵測離群值，真正的離群值容易被好的槓桿點之效果所遮掩住，而僅適用在x方向之單一離群值的情形[32]，但是作物區域試驗穩定性分析是以各品種之產量隨環境指標作迴歸所得的迴歸係數來評價品種的穩定性，該環境指標為各環境之平均收量與整個試驗的總平均值之偏差，並非固定值，因此發生在y方向的離群值，自然也會影響x方向之數值，因此僅由表3之結果，無法同時針對x及y兩向之離群值進行診斷，其實很難論定是否還有其他離群值存在。

Yates-Cochran之大麥試驗資料以PROGRESS電腦程式分析結果，各品種所估得之穩健殘差($r_i/\sigma*$)列於表4。由表中得知，只有品種2在環境1,2的$r_i/\sigma*$均甚大於其切斷點2.50，表示該兩個觀測值為迴歸離群值。倘繼之以$r_i/\sigma*$為基礎，進行加權分析，則該兩觀測值之加權值為0，其餘正常資料之加權值則為1，如表4所示，於是所得之RLS估計結果列於表5。比較表5之LS與RLS的結果差異，經過離群值的加權矯正後，品種2的直線配合滿意程度(決定係數R^2值)提高，殘差均方顯著降低，且其迴歸係數估值由0.99變為0.81，不過此迴歸係數估計結果上的差異程度，就穩定性分析的立場而言，尚不致於影響穩定性評價的結果，即品種2仍然是一個表現相當穩定的品種，因此可以推知，該兩個離群值屬於較溫和的離群值。由上述結果顯示，利用穩健殘差作為離群值診斷工具，不但可同時偵測出多個離群值，且再藉由檢視RLS與LS估值間的差異程度，可判斷此等離群值之影響能力，但由此尚無法確知離群值是x方向的槓桿點，抑或是y方向的垂直離群值。基此，再以MINVOL電腦程式求得各品種在所有環境的穩健距離(RD_i)，如表6所示，所有資料點均未超過其切斷點，顯示沒有槓桿點存在。倘以穩健殘差對穩健距離作圖，更可明確地說明出品種2的資料型式，如圖2所示，品種2在環境3,4,5,6之資料為正常的資料，而在環境1,2之資料則同屬於垂直離群值。不過在統計上，由上述RLS與LS估值之差異比較（表5），多少可以確知含及不含該等離群值對穩定性分析的影響並不嚴重，此意謂著除非有其它生物學上的原因，否則保留或刪除的影響不大。

由於穩定性分析是以各品種之產量隨環境指標作迴歸所得的迴歸係數來評價品種的穩定性，該環境指標是以各環境之平均收量與整個試驗的總平均值之偏差，作為該一試驗的環境效應估值，因此在穩定性迴歸模式中的自變數x，即環境指標估值，為觀測值y的線性組合，並非固定值，因而若有發生

表4. 大麥資料根據LMS之強健殘差($r/\sigma*$)及加權 (W)

Table 4. Robust residuals ($r/\sigma*$) and weights (W) based on the LMS for the barley data

Line	1		2		3		4		5	
	$r/\sigma*$[z]	W	$r/\sigma*$	W	$r/\sigma*$	W	$r/\sigma*$	W	$r/\sigma*$	W
1	-0.32	1	<u>8.05</u>	0	0.41	1	-1.13	1	-1.36	1
2	0.32	1	<u>9.35</u>	0	0.13	1	-0.21	1	0.03	1
3	-0.19	1	0.89	1	-0.94	1	0.30	1	0.18	1
4	1.48	1	-0.46	1	0.13	1	-0.21	1	-1.44	1
5	1.22	1	-0.89	1	-0.41	1	-1.57	1	-0.18	1
6	0.32	1	-0.46	1	-1.66	1	-0.30	1	0.03	1

[z] The cut-off value for $r/\sigma*$ is 2.50.

表 5. 大麥資料根據LS及RLS之迴歸比較

Table 5. Comparison of regression based on LS and RLS for the barley data

Line	Intercept		Slope		Residual mean squares		R^2	
	LS	RLS	LS	RLS	LS	RLS	LS	RLS
1	188.8	188.8	0.84**	0.84**	175.6	175.6	0.90	0.90
2	182.3	172.9	0.99**	0.81**	198.5	6.5	0.91	0.99
3	198.4	198.4	0.95**	0.95**	207.1	207.1	0.90	0.90
4	236.4	236.4	1.61**	1.61**	339.7	339.7	0.94	0.94
5	205.1	205.1	0.61**	0.61**	136.7	136.7	0.85	0.85
χ^2					0.9	9.7*		

*, **: Significant at 5% and 1% level, respectively.

表 6. 大麥資料根據MVE之強健距離(RD)

Table 6. Robust distances (RD) based on MVE for the barley data

Line / Environment	1	2	3	4	5
1	0.22[z]	2.43	0.25	0.21	1.82
2	1.20	2.41	0.25	1.25	0.25
3	0.22	0.25	0.56	0.25	0.25
4	0.88	0.25	0.05	0.25	2.00
5	0.22	0.25	0.25	0.25	0.25
6	0.25	0.23	0.91	0.68	0.24

[z] The cut-off value for RD is $\sqrt{\chi^2_{2,0.975}}$ =2.72.

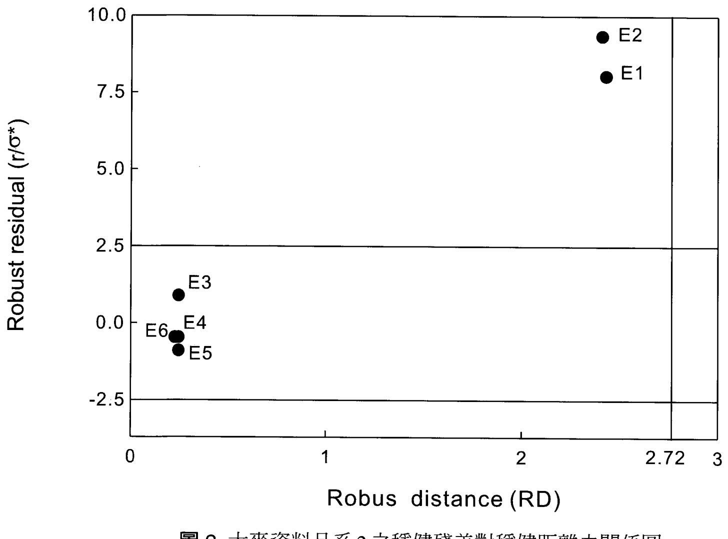

圖 2. 大麥資料品系 2 之穩健殘差對穩健距離之關係圖。

Fig. 2. Plot of robust residuals versus robust distances of line 2 for the barley data.

在y方向的離群值，自然也會影響x方向之數值。基此，以LS殘差爲基礎的傳統迴歸診斷法並不適合作爲穩定性分析的離群值診斷，其用以檢測多個離群值的效率較低[32]，且無法同時對x及y兩向之離群值進行診斷；而穩健迴歸不僅在偵測多個離群值的效率較傳統迴歸診斷法爲佳[29]，並且也能同時考慮到x及y兩向之離群值，故以穩健估值爲基礎所偵測的結果當較爲可信。而在實際應用操作上，也以穩健殘差對穩健距離之關係圖示可立即而清楚地辨識出資料的型式，以找出眞正的離群值，如再配合RLS與LS估值之差異比較，則更能判斷離群值或槓桿點對迴歸係數估值的影響程度。因此，以穩健迴歸爲主要離群值診斷法，應該較傳統迴歸診斷法更適用於作物區域試驗穩定性分析之離群值鑑定。

誌　謝

本研究承蒙行政院國家科學委員會補助經費（NSC82-0409-B055-015），及比利時 Rousseeuw, P. J. 博士提供PROGRESS及MINVOL統計分析程式，特此誌謝。

引用文獻

1. Andrews, D. F. and D. Pregibon. 1978. Finding the outliers that matter. J. R. Stat. Soc. ser. B. 40:85-93.
2. Atkinson, A. C. 1983. Diagnostic regression for shifted power transformations. Technometrics 25:23-33.
3. Atkinson, A. C. 1986. Masking unmasked. Biometrika 73:533-541.
4. Atkinson, A. C. 1987. Transformations Unmasked. Technical Report, Department of Mathematics, Imperial College, London.
5. Becker, H. C. and J. Léon. 1988. Stability analysis in plant breeding. Plant Breed. 101:1-23.
6. Beckman, R. J. and R. D. Cook. 1983. Outlier........s. Technometrics 25:119-149.
7. Belsley, D. A., E. Kuh and R. E. Welsch. 1980. Regression Diagnostics. John Wiley & Sons, New York.
8. Birch , J. B. and S. J. Fleischer. 1984. Diagnostic methods for detecting outliers in regression analysis. Environ. Entomol. 13:19-25.
9. Campbell, N. A. 1980. Robust procedures in multivariate analysis. Ⅰ. Robust covariance estimation. Appl. Stat. 29:231-237.
10. Carroll, R. J. and D. Ruppert. 1985. Transformations in regression: A robust analysis. Technometrics 27:1-12.
11. Cook, R. D. 1977. Detection of influential observation in linear regression. Technometrics 19:15-18.
12. Cook, R. D. and S. Weisberg. 1982. Residuals and Influence in Regression. Chapman & Hall, London.
13. Devlin, S. J., R. Gnanadesikan and J. R. Kettenring. 1981. Robust estimation of dispersion matrices and principal components. J. Amer. Stat. Assoc. 76:354-362.
14. Freeman, G. H. and J. M. Perkins. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. Heredity 27:15-23.
15. Hardwick, R. C. and J. T. Wood. 1972. Regression methods for studying genotype-environment interactions. Heredity 28:209-222.
16. Henderson, H. V. and P. F. Velleman. 1981. Building multiple regression models interactively. Biometrics 37:391-411.
17. Hoaglin, D. C. and R. E. Welsch. 1978. The hat matrix in regression and ANOVA. Amer. Stat. 32:17-22.
18. Hocking, R. R. and O. J. Pendleton. 1983. The regression dilemma. Commun. Stat. (Theory and Methods) 12:497-527.
19. Hogg, R. V. 1979. Statistical robustness: One view of its use in applications today. Amer. Stat. 33:108-115.
20. Holland, P. W. and R. E. Welsch. 1977. Robust regression using iteratively reweighted least squares. Commun. Stat. (Theory and Methods) 6:813-828.
21. Huber, P. J. 1964. Robust estimation of a location parameter. Ann. Math. Stat. 35:73-101.
22. Huber, P. J. 1981. Robust Statistics. John Wiley & Sons, New York.
23. John, J. A. and N. R. Draper. 1978. On testing for two outliers or one outlier in two-way tables. Technometrics 20:69-78.
24. Karnel, G. 1988. Roubust canonical correlation and correspondence analysis, unpublished manuscript, Technical University, Vienna, Dept. of Statistics.
25. Leroy, A. and P. J. Rousseeuw. 1984. PROGRESS: A Program for Robust Regression Analysis. Technical Report 201. Center for Statistics and O. R. University of Brussels. Belgium.

26. Lin, C. S., M. R. Binns, and L. P. Lefkovitch. 1986. Stability analysis: Where do we stand? Crop Sci. 26:894-900.
27. Little, J. K. 1985. Influence and a quadratic form in the Andrews-Pregibon statistic. Technometrics 27:13-15.
28. Lu, H. Y. 1992. Applicability of data transformation in regression model for stability analysis. J. Agric. Assoc. China 159:29-40. (in Chinese with English summary)
29. Pritts, M. and J. Luby. 1990. Stability indices for horticultural crops. Hortscience 25:740-745.
30. Rousseeuw, P. J. 1984. Leaste median of squares regression. J. Amer. Stat. Assoc. 79:871-880.
31. Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. In: Mathematical Statistics and Applications. Vol. B, eds. W.Grossmann, G. Pflug, I. Vincze and W. Wertz, Dordrecht: Reidel Publishing. pp283-297.
32. Rousseeuw, P. J. and A. M. Leroy. 1987. Robust Regression and Outlier Detection. John Wiley & Sons, New York.
33. Rousseeuw, P. J. and B. C. van Zomeren. 1990. Unmasking multivariate outliers and leverage points. J. Amer. Stat. Assoc. 85:633-639.
34. Rousseeuw, P. J. 1991. A diagnostic plot for regression outliers and leverage points. Compu. Stat. Data Anal. 11:127-129.
35. SAS User's Guide : Statistics. 1988. Mc:SAS Institute Inc.
36. Skroppa, T. 1984. A critical evaluation of methods available to estimate the genotype × environment interaction. Studia Forestalia Suecica 166:3-14.
37. Stevens, J. P. 1984. Outliers and influential data points in regression analysis. Psychol. Bull. 95:334-344.
38. Tietjen, G. L. 1986. The analysis and detection of outliers. In: D'Agostino, B. R. and M. A. Stephens. Goodness-of-fit Techniques. Marcel Dekker, Inc. New York and Basel. pp.497-522.
39. Velleman, P. F. and R. E. Welsch. 1981. Efficient computing of regression diagnostics. Amer. Stat. 35:234-242.
40. Virk, D. S., P. S. Virk., B. K. Manget and G. Harinarayana. 1991. Weighted regression analysis for comparing varietal adaptation. Theor. Appl. Genet. 81:559-561.
41. Westcott, B. 1986. Some methods of analysing genotype-environment interaction. Heredity 56:243-253.
42. Wright, A. J. 1976. Bias in the estimation of regression coefficients in the analysis of genotype-environmental interaction. Heredity 37:299-303.
43. Yates, F. and W. G. Cochran. 1938. The analysis of groups of experiments. J. Agric. Sci. 28:536-580.

A Demonstration of Outlier Detection on Stability Analysis of Crop Regional Trial[1]

Hsiu-Ying Lu[2]

Summary

Linear regression analysis is commonly used to assess the relative stability of varieties grown at different regions in Taiwan. However, the conventional least squares (LS) regression is susceptible to the occurrence of outliers (or unusual observations), which may have a deleterious effect on estimates of regression coefficients and on homogeneity of residual mean squares from the regression. Thus, outliers may have a significant impact on the precision of stability analysis. Many diagnostic statistics have been designed to detect outliers. They are classified into two approaches: classical regression diagnostics and recently developed robust regression. However, they are never applied in stability analysis of regional trial data. To investigate their applicability, demonstration of several diagnostics for the outlier detection using the barley data of Yates and Cochran (1938) was performed. The results revealed that the residuals from LS fits are not useful as outlier diagnostics, whereas the robust regression is useful in screening data sets and identifying outliers.

Key words：Regional trial, Stability, Outlier Diagnostics, Robust regression, Reweighted least squares regression.

---

1. Contribution No. 2023 from Taiwan Agricultural Research Institute, Council of Agriculture.
2. Senior Statistician, Department of Agronomy, TARI, Wufeng, Taichung, Taiwan, ROC.